Digital platforms and the rise of global regulation of hate speech

Digital Platforms and the Rise of Global Regulation of Hate Speech

Paolo Cavaliere
University of Edinburgh Law School

The EU Code of Conduct on hate speech requires online platforms to set standards to regulate the blocking or removal of undesirable content. The standards chosen can be analysed for four variables: the scope of protection, the form of speech, the nature of harm and the likelihood of harm. Comparing the platforms' terms of use against existing legal standards for hate speech reveals that the scope of speech that may be removed increases significantly under the Code's mechanism. Therefore, it is legitimate to consider the platforms as substantive regulators of speech. However, the Code is only the latest example in a global trend of platforms' activities affecting both the substantive regulation of speech and its governance. Meanwhile, states' authority to set standards of acceptable speech wanes.

Keywords: *Hate speech, freedom of expression, on-line intermediaries, platforms*

Forthcoming in: Cambridge International Law Journal, Volume 8 (2019) Issue 2.

1 INTRODUCTION

The Code of Conduct on countering illegal hate speech online, introduced in June 2016,[1] represents the latest attempt by the European Union (EU) to tackle the rise of illegal content online. The Code is a voluntary agreement subscribed to by a group of information technology companies (Facebook, Microsoft, Twitter and YouTube, later joined by Instagram, Google+ until its shutdown in April 2019, Snapchat, Dailymotion and Jeuxvideo) that agreed on sharing a collective responsibility in promoting freedom of expression online. These intermediaries have bound themselves to prohibit incitement to violence and hateful conduct in their community guidelines; to provide for flagging mechanisms to allow users to submit notices and set up clear and effective procedures to review any such notifications they receive; and to review the majority of them within 24 hours 'against their rules and community guidelines and where necessary national laws'.[2] After review, platforms may decide to remove or disable access to such content.[3]

The increasing centrality of platforms in setting substantive standards of acceptable speech is the focus of this work. Analysing the platforms' terms of service as de facto normative sources, as the EU Code of Conduct frames them, allows an assessment of the ways that this emerging dynamic comports with the existing legal standards set up by relevant international and regional treaties, national statutes and prominent caselaw. In the remainder of this article, the impact of the private standards included in the platforms' terms of service is assessed against current international frameworks of hate speech as found in relevant academic literature, legal provisions at the international and domestic levels, and caselaw.

The analysis is based on a framework focusing on four variables of hate speech provisions: the scope of protection, the forms of speech that the provisions seek to restrict, the nature of the harm that the different provisions seek to prevent and the causal link between the speech and the harm. Although based on prominent academic literature, the four elements feature traditionally in legislative processes and judicial reasoning. After introducing, in the next two sections, the EU Code of Conduct and the analytical framework used in this study, the article will

[1] European Union, 'Code of conduct on countering illegal hate speech online' <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en#theeucodeofconduct> accessed 13 August 2019 (EU Code of Conduct).
[2] Ibid, 2.
[3] Ibid.

illustrate how each platform addresses the variables. For each of these elements, the current state of the academic debate, legal provisions and courts' practices will be compared with the relevant guidance offered in the platforms' terms of service. The analysis will eventually inform a discussion on emerging trends in privatised regulation of speech: short-term dynamics raise concerns regarding lack of transparency, lack of accountability and lack of foreseeability of whether content posted online would cross a threshold of acceptability; long-term dynamics involve online platforms acting as speech regulators at the global level, possibly causing states' authority to set standards of acceptable speech to wane.

2 THE EU CODE OF CONDUCT ON ILLEGAL SPEECH AND THE NORMATIVE ROLE OF THE PLATFORMS' TERMS OF USE

The underpinning legal basis for the platforms' power to remove content resides in the EU Council's Framework Decision on racism and xenophobia adopted in 2008.[4] The Framework Decision outlines certain forms of conduct that amount to hate speech, such as the public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin and any such acts when carried out by the public dissemination or distribution of tracts, pictures or other material.[5] A further category of hate speech involves publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes in a manner likely to incite violence or hatred against such a group or a member thereof.[6]

In 2017, the Commission released a Communication on tackling illegal content online, urging platforms to provide clear yet detailed content policies in their terms of service.[7] The Communication clarifies that the Framework Decision does not intend to provide for the full harmonisation of hate speech laws, but rather for minimum approximation: '[t]he question of whether content is legal or illegal is governed by EU and national laws'.[8] However, the Communication continues, '[a]t the same time the online platforms' own terms of service can consider specific types of content undesirable or objectionable.'[9] This suggests two distinct categories of 'bad' speech: illegal content, defined by national and EU laws, and undesirable content as defined by the platforms themselves. The Communication explains that platforms' guidelines 'should reflect both the treatment of illegal content, and content which does not respect the platform's terms of service'.[10] As a result, platforms are to provide similar treatment to both illegal and undesirable content, as long as the terms of service provide adequate guidance in this respect. Platforms' own policies thus operate as normative bases for the removal of undesirable content akin to national laws for illegal content. The 2018 Recommendation on measures to effectively tackle illegal content online reiterates this idea, though in more nuanced terms: for instance, recital 23 requires platforms to 'provide for clarity *ex ante*, in their terms of service, on their policies on the removal or disabling of access to any content that they store, *including* illegal content.'[11]

The expectation that platforms comply with the Framework Decision and the EU Code of Conduct is one in a long series of initiatives that see private intermediaries pushed to the forefront of content regulation by national governments and international institutions. Academic

[4] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law [2008] OJ L328 (the Framework Decision).
[5] Ibid, arts 1(a) and (b).
[6] Ibid, art 1(c).
[7] Commission, 'Tacking Illegal Content Online: Towards an enhanced responsibility of online platforms' (Communication) COM (2017) 0555 final (the Communication), 16.
[8] Ibid, 16; see also 5–6.
[9] Ibid, 16.
[10] Ibid.
[11] Commission, 'Recommendation of 1.3.2018 on measures to effectively tackle illegal content online' C (2018) 1177 final, recital 23 (emphasis added).

literature has already discussed the regulatory power of intermediaries in other contexts. For instance, it has been noted that private corporations have the power to dictate contractual conditions and to control access to networks, access to sources, device interoperability and ultimately shape decisions of public policy relevance. As a result, it has been suggested that the 'danger of domination or censorship will now mostly come from private corporations instead of state powers'.[12] Specifically in the regulation of speech, the role of the UK Internet Service Providers Association's internal Code of Practice[13] in the making of key decisions on content filtering has also been observed, raising doubts about whether such forms of governance are compatible with international human rights standards.[14]

The difficulty of reconciling the growing tendency to place substantive responsibilities for content regulation on platforms with human rights standards results from two competing visions of the role of platforms in today's digital society. Platforms are often seen as public forums that facilitate discussion and participation, but also as corporate entities with the right to decide their own internal practices and standards. This fosters a narrative that limits the effectiveness of human rights standards.[15] As the case of the EU Code of Conduct demonstrates, corporate standards acquire normative capacity at the point that they complement or even supersede pre-existing norms.

3 ANALYTICAL FRAMEWORK: FOUR VARIABLES IN HATE SPEECH LAWS

The concept of hate speech is one of the most widely debated yet most elusive in legal studies. Tarlach McGonagle observes that the very term 'hate speech' does not occur literally in major legal texts, including at the international level, though it appears prominently in academic and policymaking circles where, for lack of a uniform definition, it is used to refer to a 'whole spectrum of negative discourse'.[16] In fact, Eric Heinze argues that the wording of international treaties is often excessively broad and despite recurring references to blanket bans on any kind of advocacy of hatred—such as article 20(2) of the International Covenant on Civil and Political Rights (ICCPR)[17] and article 4(a) of the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)[18]—a literal interpretation of such expansive provisions would be incompatible with the right to freedom of expression, and thus they need to be read as referring to some kinds of advocacy, but not all.[19] The qualifying elements that justify legal restrictions are

[12] Joan Barata Mir and Marco Bassini, 'Freedom of Expression in the Internet: Main Trends of the Case Law of the European Court of Human Rights' in Oreste Pollicino and Graziella Romeo (eds), *The Internet and Constitutional Law: The Protection of Fundamental Rights and Constitutional Adjudication in Europe* (Routledge, New York 2016) 71, 81.

[13] The Internet Services Providers' Association, 'Code of Practice' <https://www.ispa.org.uk/about-us/ispa-code-of-practice/> accessed 7 August 2019.

[14] Emily Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (CUP, Cambridge 2015) 134–136.

[15] Rikke F Jørgensen, 'When Private Actors Govern Human Rights' in Ben Wagner, Matthias C Kettemann and Kilian Vieth (eds), *Research Handbook on Human Rights and Digital Technology: Global Politics, Law and International Relations* (Edward Elgar, Cheltenham 2019) 346, 359–362.

[16] Tarlach McGonagle, 'Minorities and Online "Hate Speech": A Parsing of Selected Complexities' (2012) 9 Eur YB Minority Issues 419, 419–420.

[17] International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), art 20(2): 'Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.'

[18] International Convention on the Elimination of All Forms of Racial Discrimination (adopted 21 December 1965, entered into force 4 January 1969) 660 UNTS 195 (ICERD), art 4: 'States Parties condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination and, to this end, with due regard to the principles embodied in the Universal Declaration of Human Rights and the rights expressly set forth in article 5 of this Convention, inter alia: (a) Shall declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin …'.

[19] Eric Heinze, *Hate Speech and Democratic Citizenship* (OUP, Oxford 2016) 38.

a point of contention. Courts, policy- and law-makers, and academic commentators have long discussed the nature and scope of such qualifying elements, with varied and conflicting results. Building on the existing literature, this work deploys a framework for comparative analysis of norms prohibiting hate speech.

A preliminary point to note is that the European human rights framework provides for a two-tiered approach to confronting hate speech. Article 17,[20] known as the 'abuse clause', and article 10(2) of the European Convention on Human Rights (ECHR)[21] create a double-filtering mechanism, which, in the first place, restricts forms of expression deemed prima facie outside the scope of protected speech for being incompatible with the rights and freedoms provided for in the ECHR (the so-called 'guillotine effect', which excludes from the scope of article 10 categories of speech such as glorification of National Socialism, Holocaust denial and anti-Semitism on the basis of their destructive impact on other fundamental rights). Otherwise, restrictions to speech must undergo a strict scrutiny of their legality, legitimacy and necessity which also takes into account the context of the speech. Prominent literature, however, suggests a less mutually exclusive interpretation of the two tests and finds that the necessity test under article 10(2) often incorporates elements of categorical analysis similarly to the abuse clause (defined as the indirect application of article 17). Conversely, contextual circumstances of the speech are taken into consideration during the direct application of the abuse clause.[22] This reading of the caselaw of the European Court of Human Rights (ECtHR or the Court) suggests that, at the practical level, contextual factors play a fundamental role in determining whether restrictions on speech are admissible in the cases of both protected and unprotected speech.

Free speech theory supports the view that an array of contextual factors determines the legal regime applicable to speech beyond prima facie considerations about its wording or any category to which it may be assigned. In line with this analysis is Robert Post's observation that common features exist among hate speech laws, despite society-specific cultural values.[23] Two elements tend to characterise legal provisions that address hate speech: first, hate speech is defined as a statement of intrinsically negative content, typically expressing dislike, abhorrence or similar feelings; and second, the definition entails either the likelihood of causing harm or the manner in which the speech is expressed being perceived as offensive, degrading, insulting or otherwise unacceptable under relevant social norms.[24] While a statement might not amount to hate speech simply because of its content, it is often how it is presented or the harm that it causes that make otherwise protected speech fall beyond the boundaries of admissibility justifying legal restrictions.[25]

The emphasis on the effects of hate speech raises the question of the target of the negative sentiment. In this respect, Bhikhu Parekh considers that an essential feature of hate speech is that it is addressed towards a 'specified or easily identifiable' societal group (or member thereof) sharing certain characteristics.[26] Given the logical pre-eminence of the question concerning the

---

[20] Convention for the Protection of Human Rights and Fundamental Freedoms (adopted 4 November 1950, entered into force 3 September 1953) ETS 5 (ECHR), art 17: 'Nothing in this Convention may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein or at their limitation to a greater extent than is provided for in the Convention.'

[21] ECHR, art 10(2): 'The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.'

[22] Hannes Cannie and Dirk Voorhoof, 'The Abuse Clause and Freedom of Expression in the European Human Rights Convention' (2011) 29 NQHR 54, 58.

[23] Robert Post, 'Hate Speech' in Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (OUP, Oxford 2009) 123, 127.

[24] Ibid.

[25] Ibid.

[26] Bhikhu Parekh, 'Is There a Case for Banning Hate Speech?' in Michael Herz and Peter Molnar (eds), *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (CUP, New York 2012) 37, 40.

target of the speech, the framework proposed here will first analyse what groups are deemed deserving of protection (ie, the scope of protection from hate speech).

Finally, after discussing what groups are being protected, the form of the speech and the harm it causes, a final question concerning the causal link between speech and harm will be examined. The necessity for decision-makers to demonstrate that the harm is the actual result of the speech derives from the need to justify an interference with a fundamental right, such as freedom of expression, with the aim to protect another equally important value that would be put at risk by the speech. However, how to determine the causality between the speech and the harm is a difficult question that presupposes a (non-existent) generalised pattern of human behaviour and reaction to provocations, incitement and offences.[27] In fact, the likelihood that harm would ensue from the speech is framed in different ways in the US, Europe and elsewhere. The fourth variable will thus be the causal link between the speech and the harm.

4 VARIABLE I: THE SCOPE OF PROTECTION

The scope of protection refers to the subjects or groups which ought to be defended from hatred. Academic commentators have on multiple occasions remarked how the underlying rationales for regulating speech have undergone some fundamental changes through time, and the scope of protection from harmful speech has expanded or contracted accordingly. Eric Barendt notes that, during the second half of the twentieth century, media laws have become increasingly focused on the aim of preserving 'order between different groups' and less on shielding governments from criticism, which aimed instead to protect the political order from social unrest.[28] As a result, while the scope for seditious libel has been decisively reduced, group protection has come to the fore and attacks on minorities have been increasingly pushed to the margins or outside the scope of protected speech.[29] Jan Oster, however, suggests that the protection of different groups and their members coexists with the more broad-ranging protection of 'social peace in general', cohesion and public order.[30] Differences in legislative approaches can reflect this ambivalence: the scope of protection can be either defined specifically by including in relevant statutes lists of protected characteristics to identify the groups protected or left open-ended with the introduction of equal protection clauses.[31]

Despite being most common, especially among hate speech laws at the domestic level, the group-based approach has occasionally attracted criticism for its alleged under-inclusive and discriminatory nature (for 'it divides the citizenry into those who are protected and those who are not'[32]). The most recent academic debate, at least in Europe, has been widely concerned with broadening the scope of protection and making it more inclusive and wide-ranging, though it has had so far little impact on the practice of international human rights mechanisms.[33]

Far from being a matter of merely theoretical value, the prevalence of one or the other rationale affects how the other qualifying elements are operationalised and it eventually shapes courts' decisions in very practical ways. On this point, Oster again observes how in Europe a traditionally inclusive focus on preserving social peace has led courts to accept a broad range of interferences with speech in an effort to prevent attacks to societal values at large, whereas in the US, where the scope of protection is more narrowly construed and focused on specific groups, the Supreme Court has been more reluctant to restrict speech on bases other than tangible harms.[34]

---

[27] Kathleen E Mahoney, 'Hate Speech: Affirmation or Contradiction of Freedom of Expression' (1996) U Illinois L Rev 789, 797–798.
[28] Eric Barendt, *Freedom of Speech* (2nd edn, OUP, Oxford 2005) 170–171.
[29] Ibid.
[30] Jan Oster, *Media Freedom as a Fundamental Right* (CUP, Cambridge 2015) 227–228.
[31] Examples of such clauses include ICCPR, art 20(2); Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III) (UDHR) art 7: 'All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination.'
[32] Eric Heinze, 'Viewpoint Absolutism and Hate Speech' (2006) 69 MLR 543, 565.
[33] McGonagle (n 16) 422–423.
[34] Oster (n 30) 228–229.

The two aims of protecting vulnerable groups and preserving social peace/public order seem to coexist in national and international statutes, although the focus shifts from one to the other depending on historical and cultural determinants. In fact, it could be suggested that instead of being opposites, the two are rather on the same spectrum, and the wording of relevant provisions reflects the different balances struck between them.

The issue of diverging approaches to the scope of protection goes back as far as the earliest attempts to codify international treaties to tackle hate speech. In fact, debates on the opportunity to cater to cultural relativism in defining hate speech emerged as early as the *travaux préparatoires* of the Universal Declaration of Human Rights (UDHR). The UDHR does not follow a group-focused approach; rather, it includes an equal-protection clause (article 7).[35] Stephanie Farrior deduces from the drafting history of the UDHR that the clause was adopted as a limit to article 19,[36] forbidding propaganda of national, racial and religious hostility and hatred.[37] The decision to include a clause of this kind came after a proposal to include a prohibition on advocacy of racial or religious hatred and discrimination based on distinctions of race, nationality or religion was rejected amid suggestions that it would be practically unviable. The rejected wording, however, would have had the effect of establishing a direct link between advocacy of hatred and discrimination.[38] Shortly afterwards, the same connection was made explicit in international treaties such as the International Convention on the Suppression and Punishment of the Crime of Apartheid (1973), the Convention on the Elimination of All Forms of Discrimination Against Women (1979), the Declaration on the Elimination of All Forms of Intolerance and of Discrimination Based on Religion or Belief (1981), the Framework Convention for the Protection of National Minorities (1994) and equivalent statutes at the domestic level. The progressive expansion of anti-discrimination laws (which Hare describes as '[r]elated, but in many ways distinct'[39] from the historical trajectory of anti-hate speech laws) has coincidentally expanded the grounds for lawful limitations of speech on the basis of its negative impact on societal equality.

While the UDHR is arguably the first notable example of open-ended wording, the ICCPR a few years later opted for the opposite approach, although the selection of the relevant categories proved a major point of contention.[40] Eventually, a decision was reached to limit its scope to nationality, race and religion.

However, variations still exist in the approaches and wording of major general human rights treaties: article 13 of the American Convention on Human Rights includes race, colour, religion, language, and national origin.[41] The African Charter on Human and Peoples' Rights[42] and, notably, the ECHR opt instead for the open-ended wording, with equal-protection clauses such as the ECHR's prohibition of abuse of rights (article 17) and limitation of speech that infringes the rights and freedoms of others (article 10(2)) discussed above.

Stepping aside from international human rights treaties and their historical trajectories, differences in scope are most evident across national provisions. Alexander Brown identified five major categories of protected characteristics: affective states, affiliation to communities or social

[35] UDHR, art 7.
[36] UDHR, art 19: 'Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.'
[37] Stephanie Farrior, 'Molding the Matrix: The Historical and Theoretical Foundations of International Law concerning Hate Speech' (1996) 14 Berkeley J Intl L 1, 14–16.
[38] Ibid.
[39] Ivan Hare, 'Extreme Speech Under International and Regional Human Rights Standards' in Hare and Weinstein (n 23) 62, 76.
[40] Farrior (n 37) 8–10.
[41] American Convention on Human Rights (adopted 22 November 1969, entered into force 18 July 1978) (ACHR) art 13(5): 'Any propaganda for war and any advocacy of national, racial, or religious hatred that constitute incitements to lawless violence or to any other similar action against any person or group of persons on any grounds including those of race, colour, religion, language, or national origin shall be considered as offenses punishable by law.'
[42] African Charter on Human and Peoples' Rights (adopted 27 June 1981, entered into force 21 October 1986) (1982) 21 ILM 58, art 3: 'Every individual shall be equal before the law. Every individual shall be entitled to equal protection of the law.'

groups, attitudinal dispositions or beliefs, biological and physical attributes, and conducts.[43] The list-based approach is followed in the UK, where the Public Order Act 1986 prohibits speech stirring hatred on racial, religious, sexual orientation grounds. Italy prohibits incitement to discrimination on the grounds of race, ethnicity, nationality and religion.[44] The French law on the freedom of the press of 1881[45] has been amended multiple times to expand the list of protected characteristics: ethnicity, nationality, race or religion in 1972;[46] sex, sexual orientation, gender identity and disability in 2017.[47] The list of protected groups is particularly extensive in Austria and includes 'a church or religious denomination or any other group of persons defined by criteria of race, colour of skin, language, religion or ideology, nationality, descent or national or ethnic origin, sex, a disability, age or sexual orientation or a member of such a group'.[48] Most often, such lists are considered exhaustive and do not offer grounds for penalising hateful speech addressed toward other groups, as demonstrated by refusals from judicial authorities to interpret the provisions extensively or by the need to engage in legislative processes to amend the laws and include further categories. As an example of judicial refusal, English courts have been reluctant to extend the protection granted by the Public Order Act 1986 to Jews or Sikhs as victims of racial hatred to Muslims who were instead considered a religious rather than ethnic group.[49] Legislative processes include the above-mentioned successive amendments to the French law of 1881 or the recently failed attempt to add homosexuality to the list of protected characteristics in Italy.[50]

Nevertheless, some national laws are instead meant as non-exhaustive and are open to expansive interpretation. The German Criminal Code identifies a less-extensive list of groups (including national, racial, religious and ethnic groups) and adds a reference to 'segments of the population,' a seemingly open clause that could extend future applications of this provision to any minority not explicitly mentioned in the original wording.[51] The Finnish Criminal Code cites 'race, skin colour, birth status, national or ethnic origin, religion or belief, sexual orientation or disability or a comparable basis'.[52] Even more explicitly, the Romanian Criminal Code prohibits incitement to hatred or discrimination 'against a category of individuals'.[53]

Different approaches emerge among the online platforms with respect to the scope of protection. Most notably, they split almost evenly into two groups. Dailymotion, Microsoft and Snapchat do not include in their terms of use lists of protected characteristics,[54] which suggests a horizontal, open-ended approach. Facebook, Instagram, Jeuxvideo, Twitter and YouTube follow instead a list-based approach,[55] whilst Google+ took the middle way providing both a list of categories and an equal-protection clause ('any other characteristic associated with systematic

---

[43] Alexander Brown, 'The "Who?" Question in the Hate Speech Debate: Part 1: Consistency, Practical, and Formal Approaches' (2016) 29 Can J L Juris 275, 281.
[44] Legge 25 giugno 1993, n 205, art 1(A).
[45] Loi du 29 juillet 1881 sur la liberté de la presse, art 24.
[46] Loi n° 72-546 du 1er juillet 1972 relative à la lutte contre le racisme, art 1.
[47] Loi n° 2017-86 du 27 janvier 2017 relative à l'égalité et à la citoyenneté, art 170.II.1°(a).
[48] Penal Code FLG 1974/60 (as amended by FLG I 2011/103), § 283.
[49] Kay Goodall, 'Incitement to Religious Hatred: All Talk and No Substance?' (2007) 70 MLR 89, 93.
[50] Alessandro Fulloni, 'Homophobia Law Blocked' (*Italian Life*, 27 July 2011) <https://www.corriere.it/english/11_luglio_27/homophobia-law-blocked_b7509f8a-b838-11e0-a142-4db684210d8b.shtml> accessed 7 August 2019.
[51] Criminal Code in the version promulgated on 13 November 1998, Federal Law Gazette [Bundesgesetzblatt] I, 3322 (last amended by Article 3 of the Law of 2 October 2009, Federal Law Gazette I, 3214), s 130.
[52] Criminal Code (39/1889, amendments up to 766/2015 included), ch 11, s 10.
[53] Law #286 of 17 July 2009 of the Criminal Code, art 369.
[54] See Dailymotion, 'Terms of Use' <https://www.dailymotion.com/legal> accessed 13 August 2019; Microsoft, 'Microsoft Services Agreement' <https://www.microsoft.com/en-gb/servicesagreement/> accessed 28 July 2019; Snapchat, 'Snap Inc Terms of Service' <https://www.snap.com/en-GB/terms/> accessed 28 July 2019.
[55] See Facebook, '11. Hate Speech' (*Community Standards*, 2019) <https://www.facebook.com/communitystandards/hate_speech> accessed 13 August 2019; Instagram, 'Community Guidelines' (2019) <https://help.instagram.com/477434105621119> accessed 13 August 2019; Jeuxvideo, 'Charte des forums' (2019) <http://www.jeuxvideo.com/forums_charte.htm> accessed 28 July 2019; Twitter, 'The Twitter Rules' (2019) <https://help.twitter.com/en/rules-and-policies/twitter-rules> accessed 28 July 2019; YouTube, 'Community Guidelines' (2019) <https://www.youtube.com/yt/about/policies/#community-guidelines> accessed 13 August 2019.

discrimination or marginalisation').[56] The different lists present some predictable similarities. They all include disability, gender identity and sexual orientation, with the exception of Jeuxvideo, which includes sex but not gender.[57] Even more striking are some differences among the platforms' terms of use. Facebook and YouTube are the only platforms to include caste and immigration status.[58] Only Google+ and YouTube include the status of veterans.[59] YouTube also prohibits hate towards victims of major violent events and their kin and, most notably, is the only platform with no open-ended clause that left out the characteristics of serious diseases and nationality or national origin.[60] Jeuxvideo includes the category of lifestyle,[61] seemingly a catch-all expression capable of including many, but not all, situations.

Two notable tendencies emerge from this data. Most evidently, the platforms differ greatly from each other as to the quantity and quality of the guidance they give to their users. Some, such as Dailymotion and Snapchat, are extremely synthetic, whereas others, such as Facebook, Twitter and YouTube, are very detailed. Among those that opted for a list-based approach, there seems to be a fair amount of consistency, although the few differences that emerge do not seem to be justified by any specificities related to the nature or the business model of a given platform. This may raise concerns in terms of predictability of the consequences of users' posting on different services.

When compared to the scope of protection provided by national statutes and, even more, international treaties, the platforms' lists seem more extensive, as they extend protection to groups that do not enjoy it in offline speech. This tendency is consistent with the expansive approach followed by international decision-makers and a number of national jurisdictions, although it diverges from the approach of some countries, which currently maintain a more rigid approach to their lists of protected groups.

5 VARIABLE II: THE FORM OF THE SPEECH

The second qualifying element is the manner in which speech is presented. This involves more than just the style; it encompasses the whole range of human acts capable of expressing a sentiment, and thus counting as an act of speech. Judith Butler for instance considers that the US Supreme Court, when deciding in *RAV*[62] on the act of burning crosses, was in fact determining a more fundamental question: not just the contours of legitimate speech, but 'what constitutes the domain of "speech" itself … asserting its state-sanctioned linguistic power to determine what will and will not count as "speech"'.[63] Sionaidh Douglas-Scott raises a similar point noting that words can instead count as actions at times. In such cases, speech has 'transformative capacity', ie, speech that performs the act in the very moment it enunciates it: the greater the transformative capacity of words, the greater their potential to harm.[64]

Beyond the question of the thin line between acts and words, the second variable concerns also the different forms of expression that decision-makers are prepared to take into consideration when adjudicating cases of hate speech, and thus consider (at least in principle) capable of doing harm. The emphasis on certain forms of expression rather than others may depend, for instance, on how the other variables come into play in assessing the legality of the speech: Brown observes that certain forms of expression are more likely than others to cause non-material harm (such as group defamation), including the likes of false statements of facts, insults, slurs, derogatory

[56] See Google+, 'User Content and Conduct Policy' (2019) <https://www.google.com/intl/en_uk/+/policy/content.html> accessed 13 August 2019.
[57] Jeuxvideo (n 55).
[58] Facebook (n 55); YouTube (n 55); YouTube, 'Hate speech policy' (*YouTube Policies*, 2019) <https://support.google.com/youtube/answer/2801939?hl=en> accessed 13 August 2019.
[59] Google+ (n 56); YouTube (n 55); YouTube (n 58).
[60] YouTube (n 58).
[61] Jeuxvideo (n 55).
[62] *RAV v City of St Paul* 505 US 377 (1992).
[63] Judith Butler, *Excitable Speech: A Politics of the Performative* (Routledge, New York 1997) 52–53.
[64] Sionaidh Douglas-Scott, 'The Hatefulness of Protected Speech: A Comparison of the American and European Approaches' (1999) 7 Wm & Mary Bill Rts J 305, 332–333.

epithets and ridiculing.[65] Alexandra Timmer suggests a partially different interpretation, connected to the growing global efforts to tackle discrimination in different forms: as discrimination in explicit forms is increasingly pushed outside the boundaries of acceptable speech and penalised, the same rhetoric is instead furthered through more subtle forms, which as a result enter under the radar of courts.[66] Such types of expression include stereotypes that advance discriminatory beliefs: Brown describes them as expressions that 'constitute unbalanced, oversimplified, or misleading impressions of reality'.[67]

The ECtHR has accepted that restrictions of negative stereotypes, such as those that further stigmas against mental disabilities[68] or HIV,[69] can be compatible with article 10. The Court has, however, amassed quite a number of decisions in which the most disparaging forms of expression were taken into consideration, holding at least on one occasion that insults, ridicule and 'irresponsible' speech in general can amount to attacks on persons or groups.[70] The decision has been criticised as 'a shocking departure from very well-settled case law' for its apparent failure to define the contours of acceptable forms of expression with enough certainty.[71] Despite the criticism, in other decisions of a similar kind, images (in the form of a poster) were considered to amount to a religious attack,[72] while both poems, through recourse to pathos and metaphors,[73] and novels[74] were considered capable of inciting violence if taken literally.

Among the platforms' terms of use, the main emphasis is normally on text-based content, although images and videos also feature prominently, with predictable variations depending to large extent on the specific business model of each platform (for instance, the fact that YouTube operates as a video-sharing platform while Instagram is dedicated mostly to photos plays a role). In general, the focus on (more or less) direct threats and calls for violence or hatred is not different than the courts' usual approach.

Other platforms' approaches seem less common. Facebook, which separates attacks into three tiers according to grades of severity, prohibits (among other forms of expression) mockery in tier 1, '[e]xpressions of contempt or their visual equivalent including (but not limited to) … "I don't like"' in tier 2, and slurs in tier 3.[75] Instagram prohibits humour when addressed to victims or survivors of self-injury.[76] Microsoft's Code of Conduct includes language generically defined as 'offensive'.[77] Twitter's policy, definitely one of the most extensive in this respect, prohibits references to mass murder and other violent events that targeted a protected group, the spreading of fearful stereotypes about a protected category, slurs, epithets, and racist and sexist tropes.[78] YouTube's policies do not allow challenges, pranks and any other acts that may result in physical harm or emotional distress to children, as well as any display of behaviours such as consumption of hard drugs and instructional bomb-making, on the assumption that they could encourage emulation.[79]

Beyond text-based expression, there is general attention to graphic content, as could easily be expected given the nature of interaction on these platforms. Twitter gives specific attention to images and prohibits 'hateful imagery' defined as 'logos, symbols, or images whose

---

[65] Alexander Brown, *Hate Speech Law; A Philosophical Examination* (Routledge, New York 2015) 21–23.
[66] See generally Alexandra Timmer, 'Toward an Anti-Stereotyping Approach for the European Court of Human Rights' (2011) 11 H R L Rev 707.
[67] Brown (n 65) 23.
[68] *Alajos Kiss v Hungary* App no 38832/06 (ECtHR, 20 May 2010).
[69] *Kiyutin v Russia* App no 2700/10 (ECtHR, 10 March 2011).
[70] *Vejdeland and Others v Sweden* App no 1813/07 (ECtHR, 9 February 2012).
[71] Roger Kiska, 'Hate Speech: A Comparison Between the European Court of Human Rights and the United States Supreme Court Jurisprudence' (2012) 25 Regent Univ L Rev 107, 112.
[72] *Norwood v United Kingdom* App no 23131/03 (ECtHR, 16 November 2004).
[73] *Karataş v Turkey* App no 33179/96 (ECtHR, 9 July 2002).
[74] *Alinak v Turkey* App no 40287/98 (ECtHR, 29 March 2005).
[75] Facebook (n 55).
[76] Instagram (n 55).
[77] Microsoft (n 54).
[78] Twitter, 'Hateful conduct policy' (*Twitter Rules and Policies*, 2019) <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> accessed 13 August 2019.
[79] YouTube, 'Harmful or dangerous content' (*YouTube Policies*, 2019) <https://support.google.com/youtube/answer/2801964?hl=en> accessed 13 August 2019.

purpose is to promote hostility and malice'.[80] Facebook prohibits videos that show 'dying, wounded, or dead people' if very crude imagery is displayed, like dismemberment or cannibalism; videos or photos showing 'non-sexual child abuse', like kicking, forcible restraint by an adult, forcible smoking; and videos that show violent crimes like animal abuse, physical bullying, torture.[81]

The differences in the platforms' approach are more striking in this case. Extensive and detailed guidance provided by Facebook, Instagram, Twitter and YouTube can hardly be compared to that given by Dailymotion, Snapchat, and Jeuxvideo, which are practically silent on this point. It is likely that these platforms will develop a more distinct approach through their own practice and its consolidation, so the transparency of their decisions on taking down material will give clearer guidance and predictability to their users. Although there were significant variations among platforms that offer detailed guidance, one distinct trend is that they have a more expansive approach to the forms of speech than do the courts. The frequent inclusion, within prohibited content, of expressions like slurs, mockery and generically offensive content seems a marked departure from the ECtHR's approach which has consistently and decisively determined that speech could not be penalised on the sole basis of being offensive. However, the ECtHR has been prepared at times to accept limitations based on the targeting of protected groups, and there is in general a growing trend toward accepting a more diverse range of expressions as amounting to unlawful speech. Similar to the ECtHR, the platforms seem to maintain a close link between subtle forms of speech and their being addressed towards specific protected groups. What seems to be missing is that the platforms' terms of use do not engage with either the different kinds of harm or the transformative capacity of the different types of content.

Graphic representation of objectionable behaviour is generally prohibited on the grounds of its asserted potential for emulation. In general, when the European Court has accepted a limitation to a new style of expression (like poetry or fiction), this was accompanied by an explanation or qualification of the specific circumstances that would make the restriction acceptable. Platforms, in contrast, seem ready to limit several forms of expression on a rather open-ended basis. A few national jurisdictions within the EU also identify mocking and degrading speech as possibly unlawful (as discussed below), although courts traditionally perform a case-by-case assessment of the impact of each statement in the concrete circumstances. It is unclear whether platforms would have the capacity to perform such contextual assessment, which, together with the broad variety of forms of speech included in the guidance (wide enough to include even expressions of dislike in some cases), makes it seem likely that platforms will eventually take a harsher stance toward different styles of expression than most other decision-makers.

6 VARIABLE III: THE NATURE OF THE HARM

The third variable concerns the kind of harm considered relevant enough to justify intervention. There are alternative approaches in the analyses of both legal provisions and caselaw as to whether speech should be stifled only inasmuch as it triggers physical violence, or if other types of (non-physical) harm can be a sufficient basis for interferences.

The broader approach of accepting speech restrictions based on non-physical harm has sometimes been met with reluctance. Evan Simpson contends that hateful speech warrants external coercion when it is demonstrably harmful, as opposed to merely offensive.[82] Physical harm, such as bodily injury, and psychological harm, such as fear or humiliation, have the same effect in that both impair someone's capacity to pursue their own interests.[83] This offers a plausible rationale to justify restricting speech that causes non-physical harm when it has an

[80] Twitter (n 78).
[81] Facebook, '12. Violence and graphic content' (*Community Standards*, 2019) <https://www.facebook.com/communitystandards/graphic_violence> accessed 13 August 2019.
[82] Evan Simpson, 'Responsibilities for Hateful Speech' (2006) 12 Legal Theory 157, 158 and 161–162.
[83] Ibid, 163.

impairing capacity. Physical harm can be objectively observed and assessed. Emotional harm, however, must pass a 'test of reasonableness', since the victim's internal judgement plays a role in determining the magnitude of the harm.[84] Simpson calls this process the 'epistemic responsibility' of the individual and argues that, if individuals can react to moral harm in different ways, it follows that cognitive defences are in principle available, at least to some, against verbal assault.[85] This in turn justifies different legal responses, because individuals' responsibility for their personal beliefs exonerates the state from imposing its own system of judgement through legal norms.[86] This line of reasoning is relevant in that it reflects and explains the traditional prominence that law-makers and courts have accorded to the aim of protecting individuals from physical rather than psychological harm and generally equate hate speech with incitement to violence.

The emphasis on the subjective element at stake with non-physical harm is also what motivates the classic liberalist criticism that a focus on any consequence other than violence necessarily leads to arbitrary censorship. To this argument, Jeremy Waldron responds, in a way that has proven influential in the recent academic debate, that hatred can undermine the dignity and equal status in society of vulnerable minorities, which is in itself a tangible and often permanent harm.[87]

The *travaux préparatoires* of the UDHR included debates about whether explicit links between speech and violence should be made. Farrior traces them back to a draft proposed by the representative of the Soviet Union to bring 'advocacy' of hatred and discriminatory 'action' under the same framework provision of article 7.[88] Following on this proposal, the Chinese representative suggested that hate speech would be defined as speech expressly 'designed to provoke violence'.[89] While this was a very evident example of a position establishing a material connection between speech and action, its eventual rejection in the final version of article 7 offers a glimpse into the difficulty in equating hate speech with its violent consequences alone, paving the way for expansive caselaw in the years to come. In a similar fashion, early drafts of the ICCPR defined hate speech as constituting only an incitement to violence (notably building on the proposal originally tabled by the Soviet Union while drafting article 7 of the UDHR) but the definition was soon extended to include discrimination and propaganda for war.[90] The French representative proposed modifying the wording to 'incitement to violence and hatred'.[91] The debate that ensued demonstrates an intent to go beyond the narrow focus on violence. For instance, the Polish delegate suggested a causal relationship between hatred and violence and the need for a provision tackling hate speech to capture both these dimensions.[92] The delegates also debated the conjunction between the words 'hatred' and 'violence' and agreed that using either of the two would lead to different consequences. The Polish delegate remarked that using the disjunctive would help to stress that hatred, discrimination and violence amount to three distinct categories of harm, and the Soviet delegate agreed that speech that does not directly cause violence should be condemned.[93] Other positions, however, suggested a different understanding; the delegate from the Philippines connected article 20 to 'the right to life and the right to live in peace with one's neighbours', seemingly suggesting an emphasis on physical harm.[94]

It is not surprising that the differences between the US and European approaches emerge as the most striking. The US Supreme Court has famously embraced a more protective approach to free speech; it tends to warrant restrictions only in case of incitement to violence, as opposed

[84] Ibid, 162–164.
[85] Ibid, 164–165.
[86] Ibid.
[87] Jeremy Waldron, *The Harm in Hate Speech* (Harvard UP, Cambridge 2012) 34–41.
[88] Farrior (n 37) 15–16.
[89] Ibid, 16.
[90] Ibid, 21–22.
[91] Ibid, 23.
[92] Ibid, 25.
[93] Ibid, 26.
[94] Ibid, 36.

to hatred. Influential precedents such as *Brandenburg*,[95] *Skokie*,[96] *RAV*[97] and *Black*[98] all affirmed the predominance of the 'incitement to violence' standard over any emotional reactions that the speech could stir among both targeted and non-targeted audiences.[99] However, the dichotomy between material and non-material harm can also be framed in different terms. Surveying the US Supreme Court's caselaw, Frederick Schauer distinguishes between third-party harm, based on the dynamic of advocacy (speech that causes harm because it incites the listeners to attack or otherwise wrong third parties), and second-party harm (speech that causes harm directly to the listeners).[100] A fundamental feature of this reconstruction is that the nature of second-party harms is intrinsically different and wider than the 'incitement to violence' standard, remarkably including emotional distress[101] ('intentional infliction of emotional distress' was the tort considered in *Snyder v Phelps*,[102] once more a case decided against the restriction).

The ECtHR has often considered incitement to violence and to hatred cumulatively; the language used by the Court has been interpreted as suggesting that the two notions, although distinct in theory, make little difference in practice. As a result, the Court has used the same reasoning about the two issues, just removing the element of incitement to violence from the picture when hatred is at stake.[103] Notable cases in which the Court has upheld limitations based on harm to dignity include *Erbakan*[104] (in which the Court asserted that equal dignity of all human beings is foundational to democracy and it may be necessary to restrict speech that, by promoting intolerance, undermines it) and *Leroy*[105] (in which the Court found that a cartoon satirising the 9/11 attacks published just days after the event would harm the dignity of the victims) among others. In *Perinçek*[106] the Court found that article 8 of the ECHR protects personal dignity and this needs to be balanced against freedom of expression.

In a few instances, hate speech laws across Europe have reflected this approach. Explicit acknowledgements of dehumanisation, denigration or degradation can be found, for instance, in the Austrian Penal Code with its reference to 'human dignity'.[107] A similar emphasis on non-material harm is in consolidated interpretations of article 5 of the German Constitution, which is commonly read as to include the value of human dignity and to deny protection to speech that portrays individuals or groups as of lesser status in society, and since 1994 prohibits the denial of the Holocaust.[108] The Danish Criminal Code prohibits statements that 'threaten, insult or degrade' protected groups,[109] while the Finnish Criminal Code prohibits the expression of an opinion or other message in which 'a certain group is threatened, defamed or insulted',[110] the Icelandic Criminal Code punishes mockery, defamation and denigration,[111] and group defamation and public insult are also prohibited under the Polish[112] and Portuguese[113] penal codes. On the contrary, other national provisions across Europe are silent on the point, which may be understood as excluding the possibility of interpreting the harm element extensively. For instance, the Hungarian courts, including those of first instance, have regularly adopted an approach close to

[95] *Brandenburg v Ohio* 395 US 444 (1969).
[96] *National Socialist Party of America v Village of Skokie* 432 US 43 (1977).
[97] *RAV v City of St Paul* 505 US 377 (1992).
[98] *Virginia v Black* 538 US 343 (2003).
[99] Michael Rosenfeld, 'Hate Speech in Constitutional Jurisprudence' in Herz and Molnar (n 26) 242, 247–259.
[100] Frederick Schauer, 'Harm(s) and the First Amendment' (2012) 2011 Sup Ct Rev 81, 100–102.
[101] Ibid.
[102] *Snyder v Phelps* 562 US 443 (2011).
[103] Mario Oetheimer, 'Protecting Freedom of Expression: The Challenge of Hate Speech in the European Court of Human Rights Case Law' (2009) 17 Cardozo J Intl & Comp L 427, 435–438.
[104] *Erbakan v Turkey* App no 59405/00 (ECtHR, 6 July 2006).
[105] *Leroy v France* App no 36109/03 (ECtHR, 2 October 2008).
[106] *Perinçek v Switzerland* App no 27510/08 (ECtHR, 15 January 2015).
[107] Penal Code FLG 1974/60 (as amended by FLG I 2011/103), § 283(2).
[108] Douglas-Scott (n 64) 322–323.
[109] Criminal Code, Order no 909 of 27 September 2005, as amended by Act nos 1389 and 1400 of 21 December 2005, § 266 b.
[110] Criminal Code (39/1889, amendments up to 766/2015 included), ch 11, s 10.
[111] General Penal Code 1940 no 19 (12 February), art 223a.
[112] Penal Code, Act of 6 June 1997, art 257.
[113] Penal Code, art 240.

the American 'clear and present danger' test and refused to apply criminal sanctions to cases of group libel, including those of evident nastiness like explicit glorification of the Holocaust, and have accepted limitations only in cases of explicit incitement to violence.[114]

A similar duality of approaches emerges across the different platforms. All of them include references that point decisively (despite slight variations in wording) towards forbidding speech that causes violence against others. At the other end of the spectrum, Jeuxvideo also prohibits content that is offensive to 'human dignity'[115]—in what closely recalls Waldron's argument. With a stunningly vague expression, Jeuxvideo however goes as far as prohibiting content that is generically offensive in nature,[116] without a further specification of the type of harm that this would cause. In many other cases, it is possible to read between the lines and identify references to types of behaviour that would necessarily cause harm other than physical injury. Similar to the language of 'human dignity' are Twitter's prohibition of content that dehumanises, degrades or reinforces negative stereotypes about a protected category; [117] Facebook's definition of hate speech as 'violent or dehumanising speech'; [118] YouTube's prohibition of 'dehumanising' comparisons of groups or individuals with 'animals, insects, pests, disease, or any other non-human entity', 'stereotypes that incite or promote hatred' and claims of the physical or mental inferiority of individuals or groups.[119]

Walking along an imaginary line towards more tangible forms of harm, there are examples of still immaterial threats that can still manifest themselves in more tangible forms of emotional or psychological distress. One example comes from Instagram, which prohibits—alongside the possibly even more tangible blackmailing and harassment—the shaming of other users.[120] YouTube takes specific concern for its users' mental wellbeing, prohibiting content that makes victims believe that they are in physical danger or, in the case of children, causes emotional distress.[121] This last approach to child protection is similar to Facebook's, which explicitly prohibits bullying and harassment of users between the ages of 13 and 18.[122]

In another step forward, Instagram acknowledges the possibility of other categories of harm definitely leaning more toward the material end of the spectrum: threats of theft, vandalism, and financial harm; glorification of self-injury also is banned when addressed to victims or survivors.[123] In a similar manner, Facebook aims to remove content that negatively targets victims or survivors of self-injury or suicide, or which encourages such conduct.[124]

Other platforms employ vague language, such as Snapchat's Terms of Service, which acknowledge graphic violence, threats, hate speech or incitement to violence;[125] the inclusion of both hate speech and incitement to violence suggests that the former implies non-violent threats. A similar conclusion could be made about Microsoft's Code of Conduct, which requires users to refrain from either 'communicating hate speech or advocating violence against others'.[126] Even vaguer is Dailymotion's terms of use on prohibited content, which amounts to 'dangerous or illegal acts … including but not limited to incitement to violence',[127] evidently encompassing more than physical violence but with no guidance whatsoever on the boundaries of this definition. YouTube provides an explicit mention that a alleging the superiority of a group would violate its

[114] Michael Rosenfeld and András Sajó, 'Spreading Liberal Constitutionalism: An Inquiry into the Fate of Free Speech Rights in New Democracies' in Sujit Choudhry (ed), *The Migration of Constitutional Ideas* (CUP, Cambridge 2007) 142, 161–164.
[115] Jeuxvideo (n 55).
[116] The original text in French refers to '*messages ... au contenu choquant*'.
[117] Twitter (n 78).
[118] Facebook (n 55).
[119] YouTube (n 58).
[120] Instagram (n 55).
[121] Youtube (n 78).
[122] Facebook, '9. Bullying and harassment' (*Community Standards*, 2019) <https://www.facebook.com/communitystandards/bullying> accessed 13 August 2019.
[123] Instagram (n 55).
[124] Facebook, '6. Suicide and self-injury' (*Community Standards*, 2019) <https://www.facebook.com/communitystandards/suicide_self_injury_violence> accessed 13 August 2019.
[125] Snapchat (n 54).
[126] Microsoft (n 54).
[127] Dailymotion (n 54).

policy if meant to 'justify violence, discrimination, segregation, or exclusion', seemingly equating the different types of harm,[128] while a FAQ update from June 2019 clarified that the policy would go as far as removing any such claim of superiority 'even if it does not explicitly call for violence'.[129]

This analysis reveals that in this case platforms' approaches vary to an even greater extent. Facebook is apparently the platform that acknowledges the broadest variety of possible types of harm, from physical injuries to dehumanisation. Dailymotion and YouTube provide examples of explicit acknowledgment of non-physical harms, while most other platforms' terms of service are not as explicit. Nonetheless, non-material harms are widely acknowledged on average, more akin to the European approach than the US; while the incitement standard is a recurring element, a focus on second-party harm and, consequently, on non-material consequences is largely predominant. The generally strong focus on psychological harm seems most likely due to the large number of users of young and impressionable age. As already observed in discussing the other variables, some platforms go to greater lengths than others in detailing the kinds of harm that warrant blocking, while others use language that seems too vague to offer thorough guidance to users.

7 VARIABLE IV: THE LIKELIHOOD OF HARM

The fourth variable focuses on how close the link between the speech and its harmful effects (of whatever kind) needs to be in order to justify restrictions. The idea of 'incitement' lies at the very foundation of the concept of hate speech, yet what amounts to incitement in practice remains an elusive question. Comparative analysis reveals different answers, in different jurisdictions, to the question as to how far removed or speculative a reason can be to justify a lawful restriction.

The *travaux préparatoires* of the ICCPR once more reveal split positions on the issue. The Chilean delegate suggested opting for a 'preventive' approach and allowing authorities to restrict speech that could give rise to 'very real danger in the longer run'.[130] This suggestion however, like other broad conceptions of hate speech, raised fears that it would offer too much opportunity for government abuse. It was eventually rejected in favour of expressions easier to interpret and define.[131]

The expectation that restrictions be based on strict casual links between the speech and its harm can be traced back to the 'clear and present danger' test from the US Supreme Court's 1919 decision in *Schenk*, which frames the test as an endeavour to asses 'whether the words used are used in such circumstances and are of such a nature as to create a "clear and present danger" that they will bring about the substantive evils that Congress has a right to prevent'.[132] This test evidently focuses on both the circumstances of the speech and a discourse analysis of its content (and although it was not further explored in this decision, it also seems to include an assumption that the harm would be 'substantive'). It has been noted that the test has been applied with increasing strictness in the following decades,[133] culminating in even stricter iterations, such as the 'fighting words' doctrine enunciated in *Chaplinsky* with an explicit focus on expressions that 'by their very utterance inflict injury or tend to incite an immediate breach of the peace',[134] and the 'true threats' doctrine focused on direct threats to 'commit an act of unlawful violence' in *Black*.[135] Both these doctrines, however, have been interpreted by commentators as exceptions to the prevalent 'clear and present danger' test.[136]

---

[128] YouTube (n 58).
[129] See YouTube, 'FAQs: Update to YouTube's Hate Speech Policies' < https://support.google.com/youtube/thread/7467669?hl=en> accessed 19 August 2019.
[130] Farrior (n 37) 25.
[131] Ibid 25–30.
[132] *Schenck v United States* 249 US 47 (1919), 52.
[133] Douglas-Scott (n 64) 315–317.
[134] *Chaplinsky v New Hampshire*, 315 US 568 (1942), 572.
[135] *Virginia v Black* 538 US 343 (2003), 356.
[136] Kiska (n 71) 142–143.

In fact, despite its historical prominence, the 'clear and present danger' has received tough criticism from academic voices: for instance, it has been described as inconsistent and unfit to capture all nuances of hate propaganda, gender-biased and exclusionary discourses.[137] With regard to the inconsistent use of the test, *Debs* reveals that the US Supreme Court, in the same year as *Schenk*, was prepared to ban speech for its 'tendency and reasonably probable effect' to cause harm,[138] as opposed to a tighter and more explicit connection between the speech and its effects. In *Gitlow*[139] and later in *Dennis*[140] the Court accepted that the test would apply more loosely in circumstances where the severity of the danger at stake would discount the unlikeliness of its happening. Although *Brandenburg*[141] is commonly understood to have finally reinstated the principle that courts should be concerned only with what amounts to direct incitement, it still lends itself to be interpreted as prohibiting both imminent and future threats, as well as both direct and indirect incitement.[142]

In rather similar terms, in the European context, Antoine Buyse distinguishes between a consequentialist approach, where a concrete potential of violence resulting from the speech is required to suppress it, and a non-determinist approach, where a looser causation link and the mere probability of harm are accepted as enough.[143] The ECtHR has swung between the two. It has allowed a pre-emptive approach based on the consideration that state intervention once violence has occurred may be late and inadequate,[144] and it has accepted a broader understanding of hate speech not necessarily limited to immediate calls for violence, as opposed to generic advocacy.[145] At other times, it has rejected the legality of restrictions because of the lack of an imminent danger of a communist coup[146] or political unrest.[147] This lack of consistency fundamentally rests on an understanding of incitement to violence and to hatred as parts of the same continuum, which in turn is entangled with the other variable regarding the nature of the harm. The Court seems generally more prepared to take a stricter consequentialist approach (and therefore to grant a wider margin of appreciation to national authorities) when the speech is more likely to result in physical violence.[148]

Commentators have noted the ambivalent approaches of both the ECtHR and the US Supreme Court about applying more or less stringent tests of causality and the different conclusions that are reached. To some,[149] the US Supreme Court and the ECtHR have been heading in opposite directions, with the US Supreme Court abandoning the 'bad tendency' test in favour of a more stringent approach while the ECtHR, on the opposite, has substantially accepted it in decisions such as *Féret*[150] and *Le Pen*.[151] Others have observed that the ECtHR has undergone an 'important evolution' from *Güzel*[152] to *Erbakan*,[153] when it shifted from accepting a 'potential risk' to peace and democracy as enough to justify a restriction, to finding the lack of demonstrable 'actual risk' and 'imminent danger' as evidence of a violation of article 10.[154] With specific regard to racist speech, it has been suggested that the US Supreme Court and ECtHR had

[137] Mahoney (n 27) 101.
[138] *Debs v United States* 249 US 211 (1919), 216.
[139] *Gitlow v New York* 268 US 652 (1925).
[140] *Dennis v United States* 341 US 494 (1951).
[141] *Brandenburg v Ohio* 395 US 444 (1969).
[142] David G Barnum, 'The Clear and Present Danger Test in Anglo-American and European Law' (2006) 7 San Diego Intl L J 263, 278–280.
[143] Antoine Buyse, 'Dangerous Expressions: The ECHR, Violence and Free Speech' (2014) 63 ICLQ 491, 492.
[144] *Vona v Hungary* App no 35943/10 (ECtHR, 9 July 2013).
[145] *Féret v Belgium* App no 15615/07 (ECtHR, 16 July 2009).
[146] *Vajnai v Hungary* App no 33629/06 (ECtHR, 8 July 2008).
[147] *Erbakan v Turkey* App no 59405/00 (ECtHR, 6 July 2006).
[148] Buyse (n 140), 493–494 and 501–502.
[149] See generally Stefan Sottiaux, '"Bad Tendencies" in the ECtHR's "Hate Speech" Jurisprudence' (2011) 7 EuConst 40.
[150] *Féret v Belgium* App no 15615/07 (ECtHR, 16 July 2009).
[151] *Le Pen v France* App no 18788/09 (ECtHR, 20 April 2010).
[152] *Güzel v Turkey (no 1)* App no 54479/00 (ECtHR, 4 April 2006).
[153] *Erbakan v Turkey* App no 59405/00 (ECtHR, 6 July 2006).
[154] Oetheimer (n 103) 441–442.

been similar until the 1960s and then markedly diverged from the 1990s with the increased propensity of the US Supreme Court to protect speech.[155]

These oscillations demonstrate how both the consequentialist and the non-determinist approaches are ultimately acceptable under free speech theory and applied, with a certain degree of inconsistency, in different historical phases by these two courts. While the US Supreme Court seems to veer more decisively towards a consequentialist approach and the ECtHR takes a more multifaceted stance, the comparative analysis reveals a lack of clear-cut direction on both sides, with a readiness to loosen the test when severe harm is at risk emerging as a point in common.

Explicit threats and incitement are prohibited on all platforms: Jeuxvideo (incitement to hatred[156]), Microsoft (whose Code of Conduct includes references to advocacy of violence[157]), Twitter ('You may not promote violence against or directly attack or threaten other people', recites the Hateful conduct policy[158]), YouTube (which refers to hate speech as content that promotes either violence or hatred[159]). The notion of incitement is also acknowledged in Dailymotion's Terms of Use,[160] although, as noted above, it is linked to an extremely generic definition of harm. Incitement also occurs in Snapchat's Terms of Service,[161] although only in reference to violence and not hatred. In all these examples, concepts such as incitement or promotion seem to be used in an interchangeable, yet generic, sense, with no further guidance or specification as to how closely connected the cause and effect should be.

Some cases specifically require a threat to be credible or explicit. Facebook explains that content may be removed when it amounts to 'a genuine risk of physical harm or direct threats to public safety'[162] and users are forbidden to organise 'future … activity that is intended or likely to cause harm to people'.[163] Similar wording can be found in Instagram's Community Guidelines, which state that reports of 'harm to public and personal safety' including 'threats of physical harm as well as threats of theft, vandalism, and other financial harm' will be checked for their credibility.[164]

Glorification and justification (of different wrongdoings) are also mentioned often, where the causal link seems looser than in the case of incitement. For instance, Facebook aims to tackle 'content that glorifies violence or celebrates the suffering or humiliation of others';[165] Instagram: 'glorifying self-injury';[166] Jeuxvideo: 'apology of war crimes';[167] an older version of Google+'s Rules and conditions of use (prior to its shutdown): 'content that justifies or incites violence'.[168]

The causal link is then evidently very loose in some other cases. Twitter's Hateful Conduct Policy dedicates an entire, lengthy section to '[w]ishing, hoping or calling for serious harm on a person or group of people', which includes conducts such as '[h]oping that someone dies' or '[w]ishing for someone to fall victim to a serious accident', with no references to the likeliness that anyone would act on such wishes.[169] In a similar way, Facebook also prohibits comments revealing 'enjoyment of [people's or animals'] suffering' or humiliation.[170] A rather

---

[155] Erik Bleich, 'Freedom of Expression versus Racist Hate Speech: Explaining Differences Between High Court Regulations in the USA and Europe' (2014) 40 J Ethnic and Migration Studies 283, 284.
[156] Jeuxvideo (n 55).
[157] Microsoft (n 54).
[158] Twitter (n 78).
[159] YouTube (n 58).
[160] Dailymotion (n 54).
[161] Snapchat (n 54).
[162] Facebook, '1. Violence and incitement' (*Community Guidelines*, 2019) <https://www.facebook.com/communitystandards/credible_violence> accessed 13 August 2019.
[163] Facebook, '4. Coordinating harm' (*Community Guidelines*, 2019) <https://www.facebook.com/communitystandards/coordinating_harm> accessed 13 August 2019.
[164] Instagram (n 55).
[165] Facebook, '12. Violence and graphic content' (*Community Guidelines*, 2019) <https://www.facebook.com/communitystandards/graphic_violence> accessed 13 August 2019.
[166] Instagram (n 55).
[167] Jeuxvideo (n 55).
[168] Google+, 'Rules and conditions of use' (as at 16 March 2019) formerly available at <https://www.google.com/+/policy/content.html> accessed 16 March 2019.
[169] Twitter (n 78).
[170] Facebook (n 162).

curious stance is in Facebook's Community Standards in what seems a presumption of severity of threats: users are advised that in case of unclear intentions, the content may be removed.[171]

The platforms thus seem to replicate the ambivalent approaches of both the US and European jurisdictions, and they pay attention to different strengths of linkages, from the strongest in the form of incitement to the loosest. As opposed to both the US Supreme Court and the ECtHR, platforms seem prepared to accept loose links in the case of non-material harm, with some examples of ill-defined causal links.

8 CONCLUDING REMARKS

The analysis of the platforms' terms of use reveals a variety of different approaches to the blocking or removal of content. The first dynamic that emerges is that the category of undesirable content is broader than illegal speech. Compared to statutes and courts, platforms, on average, protect more characteristics, are more prepared to penalise low-intensity utterances such as slurs, pranks, and generically offensive content, and address the risk of non-material and particularly psychological harm more closely. If anything, their approaches seem more aligned with the European style of balancing competing interests than the American free speech absolutism. However, some of their approaches, particularly the loose connection between speech and non-material harm, seem to depart from both the European and the American tradition. As such, the normative stance taken by platforms' policies expands the scope of speech that can be restricted.

It may be that practical circumstances justify this stance. Dynamics of social interaction unfold differently online and offline. Danielle Keats Citron suggests that dynamics like anonymity, mobilisation of groups and group polarisation cause two specific results: they 'make it more likely that people will act destructively' and they 'enhance the destruction's accessibility, making it more likely to inflict harm'.[172] Both the likeliness of harm and its magnitude are thus greater in the online context. This does not, however, eliminate concerns for the impact on freedom of expression.

Over the course of just a few weeks in the Spring/Summer of 2019, Twitter has included speech that dehumanises religious groups among the types of expression banned on the platform,[173] YouTube has expanded its policy to prohibit racial and religious superiority, the representation of protected characteristics as illness or deficiency, and added caste and the status of victim of major violent events to their list of protected characteristics,[174] and Facebook has announced a tougher stance on white nationalism and separatism.[175] Several of the terms of service analysed above may well change substantially in the near future, yet while such changes could potentially be welcome for the safety they bring to the digital environment, the lack of transparency, lack of accountability and volatility raise concerns on a more systematic level. The EU Code of Conduct stresses the need for IT companies to promote transparency; [176] providing quantitative data and statistics on removal rates through their yearly reports and giving individual feedback to notifications, as it happens at present, seems far from enough to provide the average users with enough information to realistically assess the consequences of their posting on each individual platform.

---

[171] Facebook (n 55).
[172] Danielle Keats Citron, *Hate Crimes in Cyberspace* (Harvard University Press, Cambridge 2014) 57.
[173] Jonathan Shieber, 'Twitter Updates Hate Speech Rules to Include Dehumanizing Speech Around Religion' (*TechCrunch*, July 2019) <https://techcrunch.com/2019/07/09/twitter-updates-hate-speech-rules-to-include-dehumanizing-speech-around-religion/> accessed 8 August 2019.
[174] YouTube Help, 'FAQs: Update to YouTube's Hate Speech Policies' <https://support.google.com/youtube/thread/7467669?msgid=7467669> accessed 8 August 2019.
[175] Lois Beckett, 'Facebook to Ban White Nationalism and Separatism Content' (*The Guardian*, 27 March 2019) <https://www.theguardian.com/technology/2019/mar/27/facebook-white-nationalism-hate-speech-ban> accessed 8 August 2019.
[176] EU Code of Conduct (n 1), 2.

With all the evidence, this is a transitional phase and the current state of art suggests a trend, in the next few years, toward more structured and systematic responses to online hate speech: the most notable example in this sense is Facebook's Blueprint for Content Governance and Enforcement announced in November 2018, with the proposal of an Oversight Board for Content Decisions, an independent body set up to hear appeals on content decisions.[177] Even if platforms were to outsource decisions on content to independent bodies, the perception that in doing so platforms would disentangle themselves from making 'important decisions about free expression and safety on [their] own', as the Blueprint announced,[178] seems misplaced, and fails to capture the deeper dynamic in place.

Issues such as lack of transparency, lack of accountability and lack of foreseeability are very much contingent on the current features of regulation by platforms; these may be temporary issues with easy solutions and indeed, efforts like Facebook's Blueprint are evidence of an ongoing effort to address them. Other changes are instead deeper and structural, and concern this evolving dynamic from two different angles: the object of speech regulation, in other words what kind of speech is regulated, and the governance of speech, in other words how speech is regulated. Regarding the first angle, the analysis above has illustrated how the category of 'objectionable' or 'undesirable' speech is broader and more vaguely defined than hate speech; terms of service provide a basis for the removal of further content that would not be necessarily illegal offline.

Regarding the second angle, platforms are now beginning to operate in a rule-making capacity (as opposed to just adjudicatory, as most commonly perceived), and imposing their substantive standards at the global level. It should be noted that platforms are stretching the boundaries of speech that may be restricted against a background of greatly varying approaches across regional and international treaties, national statutes and caselaw. In this highly fragmented, fast-changing and malleable landscape, platforms bring in their own substantive standards in the absence of common international approaches. This study has been prompted by the EU Code of Conduct and its legal underpinnings for platforms' regulatory efforts. However, the dynamic it captures is definitely unfolding well beyond the boundaries of Europe, even in the absence of the legal underpinning offered by the Code. Similar conversations have been happening in the USA for at least a couple of years now[179] and platforms are sending strong signals that they expect their regulatory efforts to grow even further in the near future. In the public consultation on its Blueprint, Facebook expressed a firm view that the Oversight Board should have a global mandate due to concerns that a 'regionalised approach' with 'different rules for different countries' could mean a rush to the bottom in terms of free speech standards and independence from restrictive governments.[180] However, the emergence of global standards of speech beyond the nation state could likely erode spaces to cater for local, historical, cultural specificities and reduce levers for states to control the boundaries of acceptable speech.

In the academic debate, it has been noted that we might be on the brink of a transition to a 'community-based, self-regulatory model of jurisdiction' where platforms form 'part of a separate non-state actor-created sphere of prescriptive and enforcement jurisdiction that exists alongside jurisdictional structures of the nation state'.[181] However, as new platforms' powers emerge beyond territorial sovereignty, the next question then is how long it will take for states to

---

[177] Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement' (*Facebook*, 16 November 2018) <https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/?hc_location=ufi> accessed 8 August 2019.

[178] Ibid.

[179] See generally Natasha Tusikov, *Chokepoints: Global Private Regulation on the Internet* (University of California Press, Oakland 2016).

[180] Brent Harris, 'Global Feedback and Input on the Facebook Oversight Board for Content Decisions' (*Facebook Newsroom*, 27 June 2019) <https://newsroom.fb.com/news/2019/06/global-feedback-on-oversight-board/> accessed 8 August 2019.

[181] Cedric Ryngaert and Mark Zoetekouw, 'The End of Territory? The Re-Emergence of Community as a Principle of Jurisdictional Order in the Internet Era' in Uta Kohl (ed) *The Net and the Nation State: Multidisciplinary Perspectives on Internet Governance* (CUP, Cambridge 2017) 185, 193.

fight back and reconquer the power to govern the flow of information at the global level. Recent developments such as Germany's Network Enforcement Act[182] (NetzDG) and the UK's Online Harms White Paper[183] can be interpreted as examples of states' attempts to reclaim their normative power by requiring, in the German case, platforms to implement local standards of acceptable speech (the NetzDG requires platforms to take down illegal content as defined by the German Penal Code, as opposed to applying platforms' self-devised standards) or by tasking, in the British example, a government agency (as opposed to a board set up by the private sector) with overseeing the fulfilment of companies' commitments and enforcing action against them if needed. The outcome of this power struggles between the private sector and state authorities will become clear in the future; until then, as the role of platforms in governing speech becomes more substantial, it will be increasingly important that checks and balances and democratic accountability are maintained.

[182] Netzwerkdurchsetzungsgesetz vom (1 September 2017) (BGBl I S 3352).
[183] Department for Digital, Culture, Media & Sport, *Online Harms* (White Paper, CP 57, 2019).